Clasificación Automática de Cubiertas Terrestres en Imágenes Satelitales.

María José Álvarez[1], Paula Tristán[1,] José M. Massa[1] and Rubén Wainschenker[1]
[1] INTIA, Facultad de Ciencias. Exactas, Universidad Nacional del Centro de la Provincia de Bs.As., Campus Universitario, Tandil, Argentina
malvarez@alumnos.exa.unicen.edu.ar, {jmassa, ptristan, rfw}@exa.unicen.edu.ar

Abstract. Las imágenes satelitales pueden ser utilizadas para identificar las cubiertas presentes en la superficie terrestre, buscando clasificar agua, suelo desnudo, áreas edificadas, bosques, etc. El proceso de clasificación de imágenes permite la extracción de información contenida en ellas, complementando el análisis visual con la aplicación de técnicas cuantitativas para automatizar la identificación de los objetos contenidos en una escena. Por esta razón, resulta de gran importancia la utilización de la informática como herramienta para facilitar la incorporación de esta tecnología. En este contexto, se desarrolla una herramienta que permite clasificar cubiertas de manera automática a través de la aplicación de algoritmos de clasificación que no han sido evaluados exhaustivamente en imágenes satelitales. A partir de los resultados obtenidos se realiza un análisis detallado de los algoritmos bajo diferentes configuraciones iniciales.

Keywords: Imágenes satelitales, clasificación automática, cubiertas terrestres.

1 Introducción

Las imágenes satelitales pueden ser utilizadas en diversas actividades para identificar cubiertas presentes en la superficie terrestre, demostrando la necesidad de promover esta tecnología de forma que constituya un apoyo para reducir el tiempo invertido en la elaboración de investigaciones sobre la superficie terrestre.

Gracias a su formato digital es posible aplicar sobre ellas procedimientos matemáticos para la clasificación automática de grandes superficies. De esta manera, las imágenes son procesadas a través de algoritmos para asignar a cada pixel un tipo de cobertura de la tierra. Sin embargo, se debe tener en cuenta que durante el proceso de captura, la energía recibida por el sensor, proveniente de la superficie terrestre, sufre alteraciones debido a la presencia de la atmósfera entre el sol, la superficie terrestre y el sensor, es decir que éste recibe una porción de la energía reflejada por la superficie. Esto provoca que sea necesario realizar una serie de correcciones para atenuar las alteraciones y descartar errores introducidos por las mismas.

Entre los algoritmos automáticos de clasificación de imágenes es posible destacar los métodos basados en Region Growing y K-means debido a que se caracterizan por ser muy simples y que pueden ser adaptados y utilizados en diferentes ámbitos.

No obstante K-means supone como condición inicial que el número de *clústers* es conocido previamente, lo cual no necesariamente se cumple en las situaciones reales. De esta manera, los resultados finales son sensibles a las condiciones iniciales.

En este contexto, resulta de gran interés desarrollar una herramienta para categorizar cubiertas a través de algoritmos de clasificación eficientes y precisos, aplicándolos a imágenes satelitales, y cuyos parámetros puedan ser configurables. Además, los resultados provistos por estos algoritmos deben ser presentados de forma clara complementando la aplicación con herramientas que faciliten la evaluación de los mismos y la extracción de características y estadísticas de las cubiertas identificadas.

En la sección 2 se detallan los aspectos principales de los métodos de clasificación desarrollados para imágenes satelitales, junto con ejemplos de su aplicación.

En la sección 3 se describen los algoritmos de clasificación y métodos de inicialización propuestos e implementados en la herramienta desarrollada.

En la sección 4 se presentan los análisis realizados, describiendo la evaluación de los métodos de inicialización, el análisis de las clasificaciones aplicadas sobre cada banda de las imágenes satelitales individualmente y la comparación de K-Means con un clasificador básico como el basado en NDVI.

Finalmente, en la sección 5 se detallan las conclusiones obtenidas junto con las posibles extensiones de este trabajo.

2 Estado del arte

Una imagen segmentada en diversas clases puede ser más informativa que la imagen satelital original en cualquiera de las bandas espectrales [1]. De esta manera se busca que los algoritmos de clasificación trabajen con clases, para lo cual es necesario proveerles algún tipo de "conocimiento" sobre la relación entre estas clases y las firmas espectrales.

La clasificación de imágenes satelitales se ve afectada por la resolución espacial y espectral y puede realizarse a través de clasificadores por píxeles o por regiones. Los primeros consideran la reflectancia de cada píxel para asignarlo a una clase, mientras que los segundos consideran la vecindad del píxel mediante un proceso de agrupación de píxeles vecinos en regiones homogéneas, previo a la clasificación [2] [3].

El método de clasificación basado en Region Growing forma parte de la categoría de métodos de *clustering* por regiones, mientras que K-means se incluye en los métodos de agrupación por pixeles. A continuación se describen las aplicaciones de estos métodos en imágenes satelitales encontradas en la literatura con el propósito brindar una base para el desarrollo de nuevos métodos de clasificación.

En[4], Sant'Anna Bins describe un método de segmentación basado en Region Growing que demuestra buenos resultados en imágenes de bosques de la región del Amazonas y en regiones agrícolas. Indica que la desventaja de este método es que en cada iteración ocurren uniones que generan que la segmentación resultante dependa del orden de estas uniones. Ante este problema se propone un algoritmo que en cada iteración define un conjunto de subimágenes y el par más similar de las regiones adyacentes es unido a cada subimagen. Los resultados obtenidos son satisfactorios, ya

que los límites de las regiones tienen buena correspondencia con los contornos de las cubiertas en las imágenes de test.

Otro caso de aplicación de Region Growing es presentado en [2], en el cual se seleccionó una de las subregiones de la Llanura Pampeana en Argentina [5] y se utilizaron cuatro imágenes Landsat 5 con el fin de capturar las diferencias fenológicas entre distintos tipos de cobertura. En las clasificaciones resultantes se evaluaron las clases obtenidas por medio de índices utilizados en estudios de paisaje y demostraron que la clasificación por regiones produce patrones menos fragmentados que los obtenidos con las clasificaciones por píxeles, mejorando la caracterización de la cobertura de la vegetación y sus cambios en el tiempo.

Con respecto a K-Means, en [1] se propone una variante del algoritmo estándar cuyo objetivo es mostrar la efectividad de los algoritmos de *clustering* en términos cuantitativos evaluando la validez de los *clústers* con respecto al valor relativo de los mismos. Con este fin se proponen dos medidas: la concentración, por la cual ítems dentro de una misma categoría deben ser tan idénticos como sea posible y la contigüidad a través de la cual se indica que es preferible que pixeles adyacentes se encuentren en la misma categoría o clase. Se comprobó que estas medidas son propiedades casi ortogonales, con lo cual es posible realizar mejoras considerables en una de ellas con una mínima pérdida en la otra.

Por último, en 0 se presenta una variante denominada "*Filtering algorithm"*, la cual comienza almacenando los puntos en un *kd-tree* y, en cada etapa, los centros más cercanos a cada punto son computados y cada centro es movido al centroide de los vecinos asociados. Para cada nodo del árbol se desea mantener un subconjunto de centros candidatos. Luego, los candidatos son podados de acuerdo a como son propagados a los nodos hijos. La prueba final de este método involucró la segmentación de una imagen satelital de Landsat en la cual se presentaron muy buenos resultados del algoritmo propuesto.

3 Métodos Propuestos

En este trabajo se ha evaluado el comportamiento de diferentes clasificadores automáticos existentes que no se habían aplicado a imágenes multiespectrales de manera tal de poder comparar con los resultados que se obtienen en los métodos de clasificaciones basados en NDVI.
A continuación se presenta la evolución de los métodos implementados.

3.1 Clasificador de Máxima Distancia

En primer lugar, se desarrolla un clasificador no supervisado basado en entrenamiento denominado **Clasificador de Máxima Distancia** que, para cada pixel a clasificar, utiliza como vector característico la firma espectral compuesta por la reflectividad normalizada y corregida por el efecto Rayleigh.

La utilización de un clasificador entrenado implica una etapa de entrenamiento y una etapa de selección y aplicación del algoritmo de clasificación.

La primera etapa consiste en determinar las cubiertas que se utilizarán como muestras durante la clasificación. Estas clases están formadas por un conjunto de puntos que conforman distintas nube de puntos. La generación de las mismas se realiza de forma manual con el conocimiento del usuario que selecciona los puntos a incorporar a cada clase.

Cada nube de puntos se caracteriza por un vector n dimensional (donde n es el número de bandas de la imagen satelital) en las cuales se encuentra el valor de reflectancia en cada una de las bandas. Este vector representa la firma espectral de la nube de puntos y permitirá clasificar los pixeles de una región.

Posteriormente, la etapa de asignación consiste en clasificar una región asignando cada pixel a la clase que se encuentra a la mínima distancia (utilizando la distancia Euclídea), obtenida al comparar su firma espectral con la de los centroides de las clases. Este clasificador no opera de forma iterativa, sino que se trata de un clasificador de una sola pasada.

De un análisis preliminar de los resultados obtenidos se observó que cuanto menor es la cantidad de clases, se incrementa la cantidad de falsos positivos. Ante esta característica, se agregó una clase denominada "Sin Clasificar" a la cual se asignan los pixeles que se encuentran a una distancia mayor que un determinado umbral con respecto a todas las nubes de puntos. El valor umbral es configurable para facilitar la evaluación del clasificador.

3.2 Clasificador Basado en Region Growing

Luego, se propone aplicar diferentes métodos de inicialización para el entrenamiento de los clasificadores. Por esta razón se implementa el método de Region Growing, con el fin de que provea firmas espectrales de regiones homogéneas que sirvan como entrada del Clasificador de Máxima Distancia y de K-means.

Region Growing toma como semillas un conjunto de puntos seleccionados por el usuario y hace crecer las regiones considerando como condición de crecimiento que la distancia entre el centroide y un pixel se encuentre bajo un valor umbral. Se considera una vecindad de orden 4.

Cada vez que se agrega un punto a la región se recalcula su centroide considerando las firmas espectrales de los puntos que pertenecen a la misma hasta ese momento [7].

Los valores umbrales bajos generan regiones en las cuales pixeles similares no son considerados en el crecimiento de la región, mientras que si se eligen valores umbrales altos, el algoritmo genera regiones con una alta cantidad de puntos.

Dado que el resultado final depende del orden en el que se realizan las inclusiones de los pixeles, se pueden generar porciones inconexas. Este problema es aplicando operadores morfológicos a través de los cuales se suavizan las regiones.

El resultado de este clasificador es una serie de cubiertas segmentadas a partir de semillas seleccionadas por el usuario, que conservan un centroide que las identifica, el cual puede ser utilizado como vector característico de un clasificador diferente.

3.3 K-Means

K-means es un método de clasificación interesante para la evaluación sobre imágenes satelitales debido a su simpleza y a que no ha sido aplicado extensivamente en las mismas [8] [9]. El algoritmo de K-means implementado consta de cuatro etapas:

1. Inicialización: Consiste en la determinación de los centroides iniciales de las cubiertas a clasificar a través de métodos de inicialización.
2. Clasificación: En esta etapa comienza un proceso iterativo en el cual se asignan los pixeles a las cubiertas de acuerdo a la distancia a la que se encuentran de los centroides. Estas distancias son calculadas a través de una medida de similaridad.
3. Cálculo de centroides: Consiste en recalcular los vectores característicos de las cubiertas que con el fin de actualizarlos para la siguiente iteración.
4. Verificación de la Condición de Convergencia: Se verifica si la clasificación obtenida hasta el momento satisface la condición de finalización del algoritmo. La condición de convergencia puede determinarse indicando un determinado número de iteraciones, o bien cuando no se producen modificaciones entre dos iteraciones.

3.3.1 Métodos de Inicialización

La influencia de los métodos de inicialización sobre el resultado de un clasificador es un aspecto de relevancia para su análisis, por lo que se estudia el problema generando alternativas que tengan en cuenta información de las cubiertas a identificar [10] [11]. Cada método de inicialización provee el conjunto de centroides que representan a las cubiertas a identificar durante la ejecución de una clasificación.

En el presente trabajo se utilizaron los métodos de inicialización aleatorio [10] y manual. Luego, se propusieron alternativas que utilizan como semillas iniciales de una clasificación los centroides de nubes de puntos, las firmas espectrales medias de puntos pertenecientes a una cubierta y los resultados de clasificaciones previas, el cual toma como semillas iniciales de la nueva clasificación los centroides originados en una ejecución anterior.

3.4 NDVI (Normalized Difference Vegetation Index)

El Índice de Vegetación Diferencial Normalizado es una herramienta para monitorear la salud de la vegetación en un momento particular. Se basa en que los vegetales muestran una actividad significante en el infrarrojo cercano, mientras que los valores de rojo son bajos. Otras cubiertas presentan un comportamiento contrario.

Siendo L_ρ la reflectividad en las bandas necesarias para realizar el cálculo. Estas reflectividades toman valores en un rango de 0 a 1, como consecuencia, el NDVI varía entre -1 y +1.

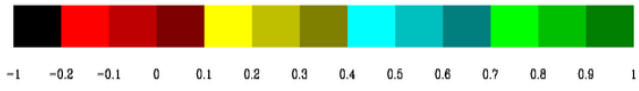

Fig. 1. Rango de valores de NDVI.

Los valores negativos y cercanos a cero indican la ausencia de vegetación. Los valores por encima de 0,2 indican presencia de vegetación, aunque puede tratarse de vegetación poco densa. A partir de valores de 0,4 se presenta vegetación con mayor cobertura. Valores de 0,6 indican áreas boscosas y por encima de 0,7 hasta 1 contienen campos de cultivo de gran vigor.

4 Análisis de Resultados

En esta sección se muestran los resultados obtenidos bajo diferentes condiciones: diferentes métodos de inicialización e integración multiespectral, los cuales se detallan en las secciones siguientes.

4.1 Evaluación de Métodos de Inicialización

Se estudió el comportamiento de los clasificadores bajo diferentes métodos de inicialización con el fin de evaluar su influencia en el resultado de las clasificaciones.

Para este fin, se eligieron muestras correspondientes a seis regiones presentes en imágenes Landsat 5 y Landsat 7 que se caracterizan por presentar superficies con agua, áreas con y sin vegetación y zonas urbanas:

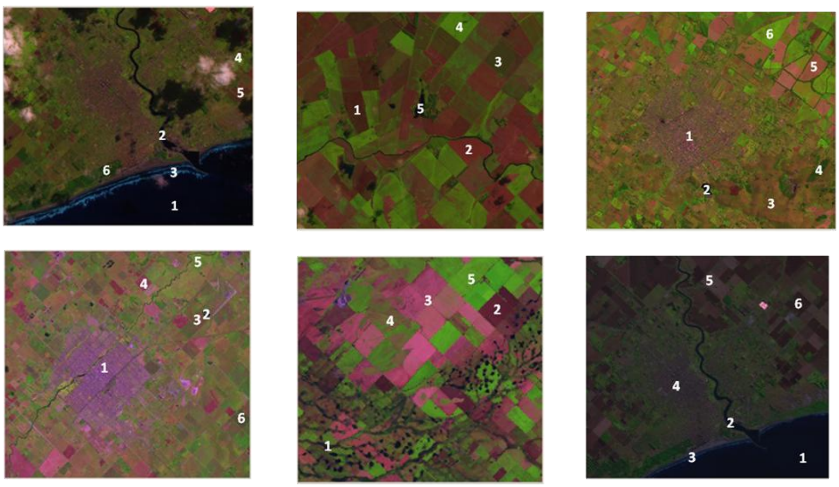

Fig. 2. Regiones bajo la Evaluación de los Métodos de Inicialización.

Estas regiones fueron clasificadas bajo los algoritmos K-means y Clasificador de Máxima Distancia y bajo los métodos de inicialización descritos en la sección 3.3.1.

De cada una de las muestras tomadas se generaron imágenes como la presentada en la Figura 3 con el fin de realizar la interpretación visual de los resultados.

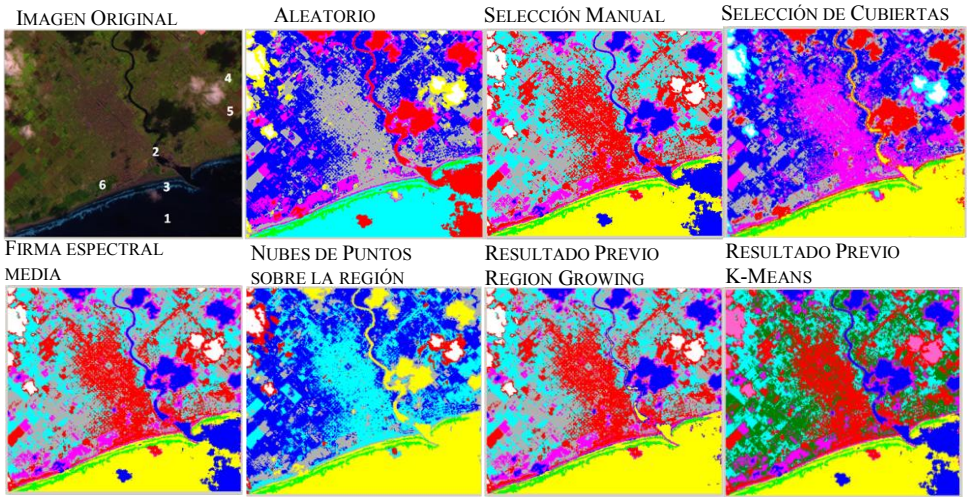

Fig. 3. Comparación visual de clasificaciones para K-means.

Además, se registró la cantidad de puntos asignada a cada una de las cubiertas identificadas, el área ocupada en la región (calculada como #pixeles * resolución espacial^2), las desviaciones estándar en cada una de las bandas, las firmas espectrales de los centroides iniciales y finales, con el fin de observar sus variaciones, y se construyeron gráficos comparativos para facilitar el análisis.

La observación de la información demostró que los métodos que presentan mayor similitud, en todas las regiones, son la selección manual, firmas espectrales medias y la utilización del resultado previo de Region Growing. Existe mayor similitud en los resultados obtenidos en K-means que en el Clasificador de Máxima Distancia.

Si se considera la utilización del resultado previo de Region Growing como una automatización de la generación de nubes de puntos, es posible comprobar las mejoras obtenidas ya que por medio de este método se obtienen áreas uniformes que presentan una gran cantidad de puntos con firmas espectrales dentro del valor umbral determinado y el usuario no interviene en la selección de los datos.

Con respecto a la utilización de resultados previos, se verificó que K-means, bajo la utilización del resultado previo de todos los métodos de inicialización, no provee mejoras en la clasificación final, ya que las mismas son realizadas en dos iteraciones y generan resultados idénticos a los originales.

4.2 Integración Multiespectral

Este análisis consiste en comparar el comportamiento de las clasificaciones aplicadas sobre cada banda de la imagen satelital con la clasificación que considera los valores de reflectancia en todas las bandas para el cálculo de las funciones de similitud y determinar cuáles presentan resultados similares a la clasificación multiespectral.

Este análisis estudia el comportamiento multiespectral de las regiones 2 y 5 bajo K-means inicializado con los métodos selección manual, firmas espectrales medias y la utilización del resultado previo de Region Growing.

Inicialmente se comparó el resultado multiespectral con las clasificaciones en cada banda:

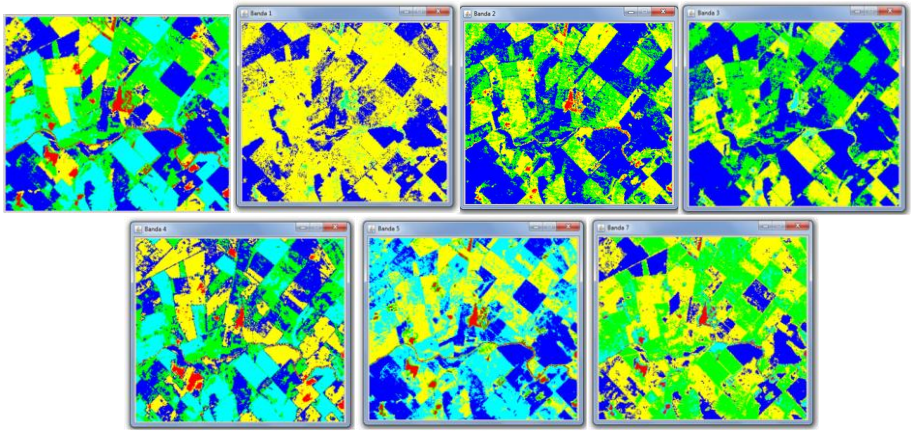

Fig. 4. Clasificación Multiespectral: comparación del Resultado Final y cada Banda.

Posteriormente se observaron los histogramas de las bandas, se aplicaron los comparadores de Diferencia e Intersección y se calcularon las matrices de confusión e indicadores asociados [13] [14] [15]. El resultado multiespectral se indica en las columnas de la matriz, mientras que las bandas se indican en las filas. Los valores en la diagonal muestran los pixeles clasificados de la misma manera, mientras que los valores fuera de la diagonal representan puntos asignados a diferentes clases.

El análisis de los histogramas demostró que en las bandas más altas se produce un aumento de la amplitud. Las bandas 4, 5 y 7 presentan los valores mayores en ambas regiones. Posteriormente, en el análisis de las diferencias e intersecciones, se verificó que en la Región 2 la banda 4 es la que presenta mayor similaridad con respecto al resultado final bajo todos los métodos de inicialización, mientras que en la Región 5 se produce en la banda 70.

Finalmente, se interpretaron los indicadores de la matriz de confusión:

- Exactitud Global: los valores mayores se presentan en las bandas 4 y 70 en las dos regiones y el rango numérico de los mismos se encuentra entre 0,61 y 0,68.
- Coeficiente de acuerdo aleatorio: Indica las probabilidades a priori de la clasificación considerando que todas las clases tienen la misma probabilidad de ser asignadas a un pixel. Dado que se identificaron 5 cubiertas, este valor es de 0,2.
- Coeficiente de Acuerdo al azar: Representa el porcentaje de similitud que se espera al azar y en ambas regiones no superó el 25%.
- Índice de Kappa: Indica la calidad de la clasificación eliminando la fracción de la clasificación debida al azar. En ambas regiones el grado de acuerdo es moderado, debido a que no supera el 0,58 en ninguna de las bandas.
- Coeficiente de Tau: En ambas regiones presenta valores entre el 50% y el 60% presentando los valores más altos en las bandas 4 y 70, respectivamente [17].

4.3 Comparación con NDVI

En esta sección se comparan los resultados de las clasificaciones obtenidas a través de K-Means con respecto a los resultados provistos al aplicar NDVI sobre dos regiones diferentes:

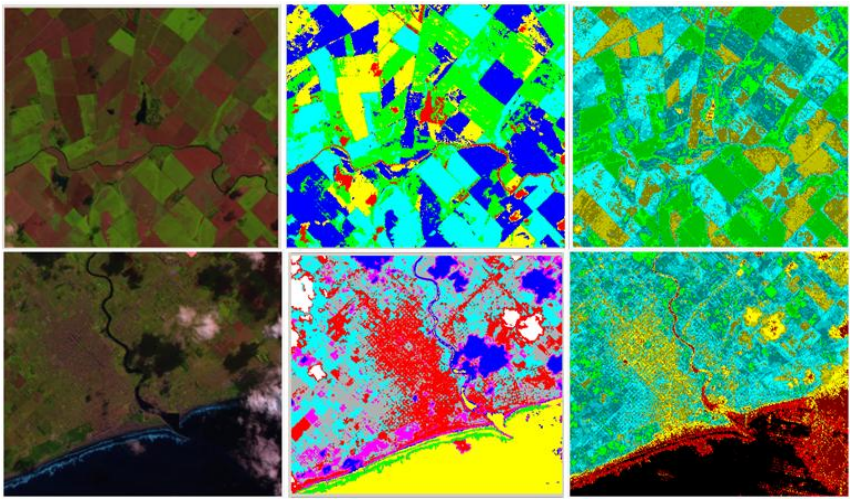

Fig. 5. Comparación K-Means (centro) vs. NDVI (derecha).

La observación de la Fig. 5 demuestra que NDVI genera aproximadamente un 40% de falsos positivos con respecto a K-Means, debido a que únicamente considera la diferencia entre dos bandas, mientras que los algoritmos de clasificación tienen en cuenta todas las bandas de la imagen satelital. Esta característica permite identificar los diferentes tipos de cubiertas presentes en una región y no solamente vegetación.

Obviamente, el NDVI permite indicar con buena precisión la salud de la vegetación, y se lo considera un indicador útil para complementar con la aplicación de clasificadores con el fin de obtener resultados de mayor precisión.

5 Conclusiones

En el presente trabajo se evaluó el comportamiento de los clasificadores implementados sobre regiones contenidas en imágenes de Landsat 5 y 7. Los análisis se centraron en la evaluación de los métodos de inicialización, el estudio multiespectral y en la comparación de K-Means con el NDVI.

La evaluación de los métodos de inicialización determinó que los que presentan mejores resultados son la selección manual, firmas espectrales medias y la utilización del resultado previo de Region Growing, mientras que los demás métodos presentan resultados con mayores diferencias en ambos clasificadores.

Posteriormente, el análisis de las clasificaciones por banda demostró que las bandas 4, 5 y 7 presentan mayores amplitudes de sus histogramas en ambos satélites y sus clasificaciones son las más semejantes al resultado multiespectral del clasificador.

Finalmente la comparación de los resultados de los clasificadores con respecto a NDVI demostró sus resultados proveen una gran cantidad de falsos positivos, concluyendo que K-Means y NDVI pueden ser utilizados complementariamente; un análisis de NDVI posterior a la clasificación permitiría mayor precisión en la clasificación de superficies con vegetación dado que este no identifica otras cubiertas.

Como resultado de este trabajo, se desarrolló una herramienta de software que permite clasificar automáticamente imágenes Landsat 5 y 7 y que, además de proveer alternativas de visualización de cubiertas clasificadas, permite la extracción de información de cada clúster, como los centroides, área ocupada, entre otros. La herramienta puede ser extendida mediante la implementación de algoritmos que no requieran la participación del usuario en la selección de los centroides iniciales.

Referencias

1. Theiler, James y Gisler, Galen. A contiguity- enhanced k-means clustering algorithm for unsupervised multispectral image segmentation. Los Alamos, USA, 1997.
2. Conde, M.C., Perelman, S.B. y Cerezo, A. Efecto de diferentes métodos de clasificación de imágenes satelitales sobre índices de paisaje. Departamento de Métodos cuantitativos y sistema de información. Facultad de Agronomía-UBA, Ciudad de Buenos Aires, 2009.
3. Quattrochi, D.A. y Pelletier, R.E. Remote sensing for Analysis of Landscape: An Introduction. Quantitative Methods in Landscape Ecology. Gardner, New York: Springer-Verlag : Ed. M. G. Turner & R. H., 1991. Págs. 51-76.
4. Sant'Anna Bins, Leonardo; García Fonseca, Leila M.; Erthal, Guaraci Jose; Mitsuo Ii, Fernando. Satellite Imagery Segmentation: a region growing approach. Brasil, 1996.
5. León, R. J. C. Geographic limits of the region, Geomorphology and geology, Regional subdivisions, Floristic aspects, Description of the vegetation. Elsevier, Amsterdam 1991.
6. Kanungo, Tapas; Netanyahu, Nathan S. ; Wu, Angela Y. ;. An Efficient k-Means Clustering Algorithm: Analysis and Implementation. IEEE Transactions On Patterns Analysis and Machine Intelligence, 2002.
7. Edman, Matt. Segmentation Using a Region Growing Algorithm. Rensselaer Polytechnic Institute, 2007.
8. Duda, R.O. y Hart, P.E. Pattern Classification and Scene Analisys. New York : Wiley, 1973.
9. Rekik, Ahmed; Zribi, An Optimal Unsupervised Satellite image Segmentation Approach Based on Pearson System and k-Means Clustering Algorithm Initialization. France, 2009.
10. Peña, J.M., Lozano, J.A. y Larrañaga, P. An empirical comparison of four initialization methods for the K-Means algorithm. E-20080 San Sebastián, Spain, 1999
11. Marina Meila, David Heckerman. An Experimental Comparison of Several Clustering and Initialization Methods. Microsoft Corporation. 1998
12. Gates, David M. Biophysical Ecology, Springer-Verlag, New York, 1980. Pág. 611.
13. Congalton, R. A Practical Look at the Sources of Confusion in Error Matrix Generation. En Photogrammetric Engineering and Remote Sensing. 1993. Págs. 641-644. Vol. 5.
14. Naeset, E. Testing for marginal homogeneity of remote sensing classification error matrices
15. Lewis, H.G y Brown, M. A generalized confusion matrix for assessing area estimates from remotely sensed data. 2001. Vol. 16.with ordered categories. 1995. Vol. 2.
16. Fleiss, JL. Measuring nominal scale agreement among many raters. Psychol Bull, 1971.
17. Zhenkul, M. Tau Coefficients for Accuracy Assessment of Classification of Remote Sensing Data. En Photogrammetric Engineering and Remote Sensing. Págs. 435-439.